

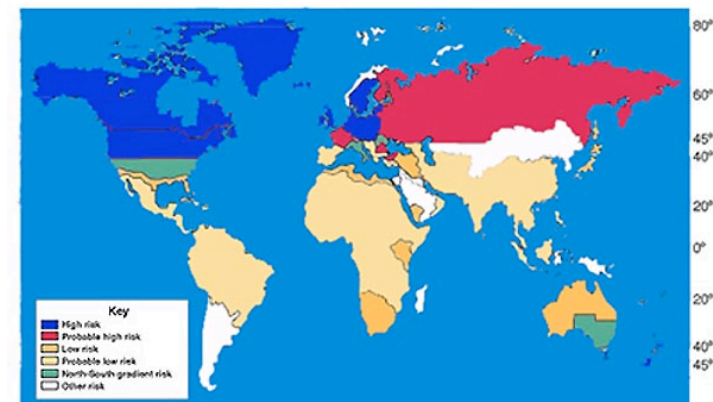
# MS Lesion Segmentation Results 3D Segmentation in the Clinic II MICCAI 2008

Martin Styner, Simon Warfield  
Olivier Commowick, JooHWi Lee  
Brian Chin, Matthew Chin, Valerie  
Jewells, Hoai-Huong Tran



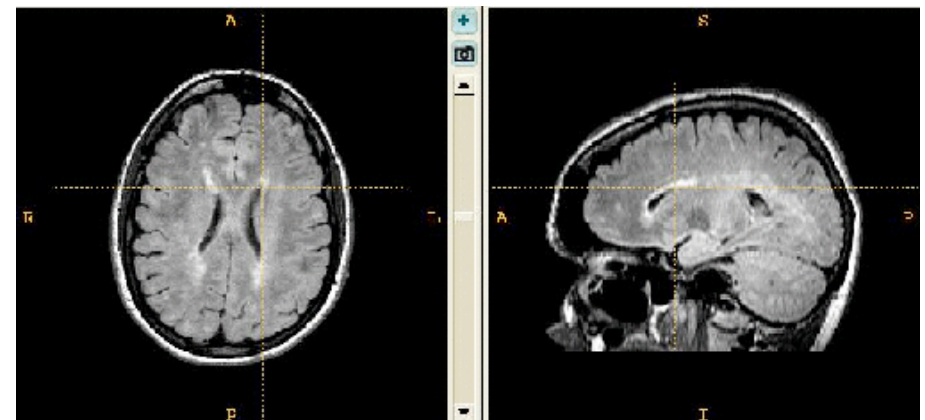
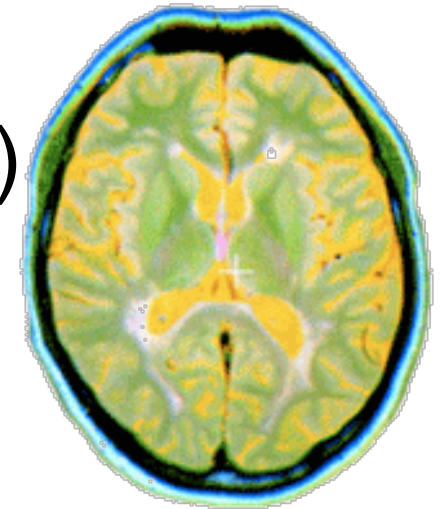
# MS Lesion Segmentation

- Multiple Sclerosis
  - Demyelinating disease
  - Auto-immune, prevalence: ~1 per 1,000
  - Inflammation in white matter
    - Destroys oligodendrocytes, myelin maintenance
    - Visible as lesions



# MS lesions in MRI

- Hyper-intense lesions (T2/FLAIR)
- Chronic/Changing lesions
- Periventricular lesions
- Expected within white matter
- FLAIR



# Competition Data

- UNC datasets (Valerie Jewells)
  - 10 training, 10 + 4 testing datasets
- Boston CHB datasets
  - 10 training, 15 + 3 testing datasets
- Subjects at various stages of disease
- Pre-workshop and onsite testing: total 53 datasets
  - High-res T1w, T2w, FLAIR, DTI-FA, DTI-MD
  - Aligned and resampled at 0.5mm isotropic
    - Original data had 0.5x0.5mm in-plane for most datasets
    - Coarse b-spline based NMI registration (Rueckert) for DTI
- Random assignment to training and testing



# Competition Data

- Reference data from 3 raters:
  - Rater CHB: all cases (on time)
  - Rater UNC-1: only UNC training cases
  - Rater UNC-2: all cases (too late)
- 2 protocols: UNC-1 and CHB, similar on paper
  - UNC-2: trained on UNC method, adapted to match CHB
- Manual segmentation
  - Time intensive job: per case 1-2 hours (UNC, InsightSNAP)
- Training on CHB-rater
- Testing on CHB & UNC-2



# Metrics/Scores

- General metrics/scores
  - Volume difference:  $\text{mm}^3$
  - Volumetric overlap (Tanimoto) error: 0 is best
  - Mean surface distance
  - True-Positive rate:  $\#(\text{Seg overlaps Ref}) / \#\text{Ref}$ 
    - 100 is best
  - False Negative rate:  $\#(\text{Seg misses Ref}) / \#\text{Seg}$ 
    - 0 is best, empty segmentation gives best score
- Volumetric too sensitive & variable, excluded
- Pre-workshop & onsite performance average



# Scores

- Scores are relative to expert variability
  - Within expert score at 90
  - Volume difference: 68%
  - Mean surface distance: 4.85mm
  - Overlap error: 75%
  - True-Positive rate: 68%
  - False Negative rate: 32%



# STAPLE

- STAPLE = Simultaneous truth and performance level estimation
- Reference Standard
  - Estimated as an optimal linear combination of rater segmentations and prior probability.
- Performance assessment
  - Characterized by sensitivity and specificity
  - Evaluated with respect to the estimated reference standard



# Measures of performance from STAPLE



- Sensitivity
  - True positive rate
  - Measures ability to segment the lesions
- Specificity
  - True negative rate
  - ability to avoid incorrect lesions
- Positive predictive value
  - Probability that a voxel labeled lesion really is a lesion
  - Ratio true positives to sum true positives + false positives
  - Good compromise combining sensitivity and specificity.
- Other summary measures can be computed too.



# The Methods

- 30 downloads, 9 submissions
- All fully automatic
  - 3 Outlier detection from classification
  - 4 intensity based classifier adapted for MS task
  - 1 Vector image joint histogram classifier
  - 1 Contextual model



# Onsite Testing Result

1 <sup>st</sup> Author	Site	Rank Final
JC Souplet	INRIA	1 / 77
P Anbeek	Utrecht	2 / 76
N Sheen	JHU	3 / 75
D Garcia	Visage	4 / 71
S Bricq	MIV	5 / 68
J Morra	UCLA	5 / 68
M Scully	MRN	7 / 66
M Prastawa	Utah	7 / 66
DJ Kroon	Twente	9 / 61



# The Winning Scoresheet



Ground Truth	UNC Rater								CHB Rater								STAPLE			
	Volume Diff. [%] Score	Avg. Dist. [mm] Score	True Pos. [%] Score	False Pos. [%] Score	Volume Diff. [%] Score	Avg. Dist. [mm] Score	True Pos. [%] Score	False Pos. [%] Score	Total	Specificity	Sensitivity	PPV								
UNC test1 Case01	56.6	92	4.8	90	51.2	81	75.7	64	36.5	95	5.3	89	53.1	82	69.2	68	82	0.9936	0.3248	0.7049
UNC test1 Case02	7.7	99	4.1	92	54.4	82	72.5	65	87.8	87	3.5	93	40.9	75	31.9	90	85	0.9949	0.1782	0.8373
UNC test1 Case03	46.2	93	1.5	97	53.5	82	49.4	80	30.5	96	1.3	97	55.9	83	40.6	85	89	0.9947	0.6529	0.8708
UNC test1 Case04	90.6	87	8.6	82	34.2	71	86.4	57	85.5	87	7.1	85	48.1	79	86.4	57	76	0.9955	0.0587	0.5161
UNC test1 Case05	53.4	92	6.0	88	52.4	81	75.2	64	5.3	99	6.4	87	78.3	96	77.9	62	84	0.9906	0.3695	0.6164
UNC test1 Case06	81.1	88	7.0	86	41.4	75	71.4	66	15.5	98	19.4	60	62.5	87	86.9	57	77	0.9932	0.1438	0.5770
UNC test1 Case07	45.6	93	2.6	95	55.7	83	58.3	74	26.2	96	6.4	87	80.0	97	71.5	66	86	0.9883	0.5412	0.6463
UNC test1 Case08	21.1	97	3.9	92	63.8	88	65.2	70	28.9	96	10.6	78	94.4	100	76.1	63	85	0.9865	0.5983	0.5587
UNC test1 Case09	157.9	77	44.0	9	33.3	70	96.6	51	263.9	61	53.0	0	0.0	51	100.0	49	46	0.9733	0.1110	0.0873
UNC test1 Case10	73.0	89	13.0	73	45.0	77	85.2	58	525.2	23	22.5	54	83.3	99	93.8	53	66	0.9847	0.6041	0.5892
CHB test1 Case01	70.3	90	3.9	92	52.0	81	27.6	93	57.6	92	2.6	95	83.9	99	35.3	88	91	0.9990	0.2075	0.8462
CHB test1 Case02	32.7	95	6.2	87	63.6	88	84.1	58	71.3	90	2.3	95	63.2	87	41.3	84	86	0.9965	0.2960	0.8138
CHB test1 Case03	48.9	93	11.6	76	50.0	80	87.7	56	75.3	89	12.2	75	33.3	70	90.4	55	74	0.9958	0.2282	0.4562
CHB test1 Case04	50.2	93	5.2	89	72.7	93	81.0	60	76.1	89	3.8	92	72.2	93	50.0	79	86	0.9992	0.1735	0.9193
CHB test1 Case05	36.8	95	8.8	82	48.1	79	92.5	53	74.0	89	3.2	93	69.6	91	71.0	66	81	0.9940	0.1492	0.5690
CHB test1 Case06	69.9	90	6.7	86	16.7	61	93.0	53	68.6	90	6.6	86	27.3	67	90.6	54	73	0.9819	0.0284	0.1001
CHB test1 Case07	62.7	91	5.5	89	48.3	79	85.6	58	77.3	89	4.8	90	55.3	83	74.0	65	80	0.9913	0.1085	0.4856
CHB test1 Case08	49.6	93	2.1	96	81.5	98	74.8	64	66.3	90	2.3	95	61.8	87	62.9	71	87	0.9964	0.3129	0.8219
CHB test1 Case09	56.3	92	3.6	92	34.9	71	69.6	67	63.1	91	2.6	95	35.8	72	55.4	76	82	0.9949	0.2781	0.7949
CHB test1 Case10	3.4	100	7.4	85	68.4	90	93.4	53	49.4	93	3.9	92	79.3	97	76.4	63	84	0.9908	0.1984	0.5123
CHB test1 Case11	16.3	98	6.0	88	47.7	79	91.4	54	72.9	89	3.6	93	51.7	81	81.7	60	80	0.9920	0.1657	0.5317
CHB test1 Case12	86.0	87	7.1	85	12.0	58	81.7	60	86.1	87	6.8	86	15.4	60	77.7	62	73	0.9872	0.0200	0.1141
CHB test1 Case13	50.0	93	5.4	89	50.0	80	85.4	58	69.3	90	4.6	91	76.2	95	64.6	70	83	0.9938	0.3095	0.7506
CHB test1 Case15	71.9	89	2.6	95	53.4	82	44.6	82	62.9	91	1.9	96	57.4	84	49.7	79	87	0.9965	0.2639	0.8947
All Average test1	55.8	92	7.4	85	49.4	80	76.2	63	86.5	87	8.2	84	57.5	84	69.0	68	80	0.9919	0.2634	0.6089
All UNC test1	63.3	91	9.5	80	48.5	79	73.6	65	110.5	84	13.5	73	59.7	85	73.4	65	78	0.9895	0.3583	0.6004
All CHB test1	50.4	93	5.9	88	50.0	80	78.0	62	69.3	90	4.4	91	55.9	83	65.8	70	82	0.9935	0.1957	0.6150
UNC test2 Case01	44.1	94	1.8	96	61.1	86	25.6	94	41.8	94	1.6	97	78.1	96	29.4	92	94	-	-	-
UNC test2 Case02	24.2	96	14.8	70	40.0	74	89.4	55	467.8	32	41.0	16	66.7	89	97.0	51	60	-	-	-
UNC test2 Case03	2.5	100	18.5	62	40.0	74	90.8	54	273.3	60	29.6	39	100.0	100	95.4	52	68	-	-	-
UNC test2 Case04	7.7	99	15.2	69	50.0	80	89.7	55	73.1	89	33.1	32	75.0	94	94.3	52	71	-	-	-
CHB test2 Case01	73.1	89	4.6	91	52.5	81	60.8	73	76.9	89	2.8	94	66.7	89	55.6	76	85	-	-	-
CHB test2 Case02	27.4	96	5.8	88	45.3	77	83.4	59	70.5	90	3.3	93	44.0	76	72.0	66	81	-	-	-
CHB test2 Case03	22.8	97	33.5	31	0.0	51	100.0	49	150.8	78	36.8	24	0.0	51	100.0	49	54	-	-	-
All Average test2	28.8	96	13.5	72	41.3	75	77.1	63	164.9	76	21.2	56	61.5	85	77.6	62	73	-	-	-
All UNC test2	19.6	97	12.6	74	47.8	79	73.9	65	214.0	69	26.3	46	79.9	95	79.0	62	73	-	-	-
All CHB test2	41.1	94	14.6	70	32.6	70	81.4	60	99.4	85	14.3	71	36.9	72	75.9	63	73	-	-	-
All Average	43.6	94	10.7	78	44.7	77	76.7	63	123.3	82	14.6	70	58.1	84	73.5	65	77	-	-	-
All UNC	41.5	94	11.1	77	48.1	79	73.7	65	162.3	76	19.9	59	69.8	90	76.2	63	75	-	-	-
All CHB	45.7	93	10.2	79	41.3	75	79.7	61	84.3	88	9.3	81	46.4	78	70.8	67	78	-	-	-



# More Results: Pre vs Onsite

1 <sup>st</sup> Author	Site	Rank Pre	Rank Onsite	Rank Final
S Bricq	MIV	1 / 82	9 / 54	5 / 68
JC Souplet	INRIA	2 / 80	2 / 73	1 / 77
N Sheen	JHU	3 / 80	3 / 69	3 / 75
D Garcia	Visage	4 / 79	5 / 62	4 / 71
P Anbeek	Utrecht	5 / 78	1 / 75	2 / 76
J Morra	UCLA	6 / 77	6 / 61	5 / 68
M Scully	MRN	7 / 67 / 73	7 / 59	7 / 66
M Prastawa	Utah	8 / 66	4 / 65	7 / 66
DJ Kroon	Twente	9 / 66	8 / 56	9 / 61



# More Results: Metrics

1 <sup>st</sup> Author	Site	Rank Final	Rank Volume	Rank Distance	Rank TPR	Rank FNR
S Bricq	MIV	68	84	53	64	72
JC Souplet	INRIA	77	88	74	81	64
N Sheen	JHU	75	79	75	80	65
D Garcia	Visage	71	83	54	72	74
P Anbeek	Utrecht	76	87	79	78	61
J Morra	UCLA	68	86	58	63	78
M Scully	MRN	66	84	52	69	59
M Prastawa	Utah	66	86	48	57	72
DJ Kroon	Twente	61	44	72	76	53



# Results: CHB vs UNC rater

1 <sup>st</sup> Author	Site	Rank CHB	Rank UNC	Rank Final
S Bricq	MIV	67	69	68
JC Souplet	INRIA	75	78	77
N Sheen	JHU	75	74	75
D Garcia	IRISA	72	70	71
P Anbeek	Utrecht	78	74	76
J Morra	UCLA	67	70	68
M Scully	MRN	65	66	66
M Prastawa	Utah	65	67	66
DJ Kroon	Twente	58	64	61



# Conclusion

- We have a winner!
- Metrics matter, but overall agreement
  - Slightly different ranking
- UNC vs CHB rater ratings differ minimally, despite only training on CHB rater
  - 2 comparable expert segmentations
  - Online competition continued after workshop
- STAPLE recomputation: Rankings

